DOCUMENT RESUME

ED 041 044 | 24 | TM 000 010

AUTHOR	Patalino, Marianne
TITLE	Rationale and Use of Content-Relevant Achievement Tests for the Evaluation of Instructional Programs.
INSTITUTION	California Univ., Los Angeles. Center for the Study of Evaluation.
SPONS AGENCY	Research and Development Centers Branch, DEL.
REPORT NO	CSE-56
BUREAU NO	BR-6-1646
PUB DATE	May 70
CONTRACT	OEC-4-6-06146-1909
NOTE	52p.
EDRS PRICE	EDRS Price MF-$0.25 HC-$2.70
DESCRIPTORS	*Achievement Tests, *Compensatory Education Programs, Course Evaluation, *Diagnostic Tests, *Disadvantaged Youth, Experimental Teaching, *Instructional Programs, Junior High School Students, Mathematics, Program Evaluation, Test Construction
IDENTIFIERS	LAMMP, *Los Angeles Model Mathematics Project

ABSTRACT

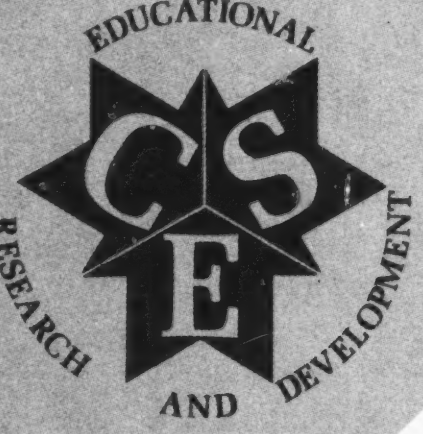
Problems in current course evaluation methods are discussed and an alternative method is described for the construction, analysis, and interpretation of a test to evaluate instructional programs. The method presented represents a different approach to the traditional overreliance on standardized achievement tests and the total scores they provide. This method led to a content-relevant, change-sensitive test which was applied to the evaluation of the Los Angeles Model Mathematics Project (LAMMP), a compensatory program for disadvantaged junior high school students achieving one or more years below grade level in mathematics. Pretest and posttest results are analyzed at the item level and interpreted in terms of specific classroom activities. This type of approach may contribute to the effectiveness of course evaluation and provide currently needed alternatives to present practices in course evaluation. Two forms of the LAMMP Diagnostic Test and a bibliography are included. (RF)

ED041044

BR 6-1646
PA 24
TM

CENTER FOR THE
STUDY OF
EVALUATION

UCLA
Graduate School
of Education
Los Angeles, California



RATIONALE AND USE OF CONTENT-RELEVANT ACHIEVEMENT TESTS FOR THE EVALUATION OF INSTRUCTIONAL PROGRAMS

TM000010

Marianne Patalino

CSE Report No. 56
May, 1970

CENTER FOR THE STUDY OF EVALUATION

Marvin C. Alkin, Director
Publications Committee:
James W. Trent, Chairman
Theodore R. Husek
Sherman J. Pearl
Audrey Schwartz

UCLA Graduate School of Education

The CENTER FOR THE STUDY OF EVALUATION is one of nine centers for educational research and development sponsored by the United States Department of Health, Education and Welfare, Office of Education. The research and development reported herein was performed pursuant to a contract with the U.S.O.E. under the provisions of the Cooperative Research Program.

Established at UCLA in June, 1966, CSE is devoted exclusively to finding new theories and methods of analyzing educational systems and programs and gauging their effects.

The Center serves its unique functions with an inter-disciplinary staff whose specialties combine for a broad, versatile approach to the complex problems of evaluation. Study projects are conducted in three major program areas: Evaluation of Instructional Programs, Evaluation of Educational Systems, and Evaluation Theory and Methodology.

This publication is one of many produced by the Center toward its goals. Information on CSE and its publications may be obtained by writing:

Office of Dissemination
Center for the Study of Evaluation
UCLA Graduate School of Education
Los Angeles, California 90024

ED041044

U.S. DEPARTMENT OF HEALTH, EDUCATION
& WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

RATIONALE AND USE OF CONTENT-RELEVANT ACHIEVEMENT TESTS FOR THE EVALUATION OF INSTRUCTIONAL PROGRAMS

Marianne Patalino

CSE Report No. 56
May, 1970

Center for the Study of Evaluation
UCLA Graduate School of Education
Los Angeles, California

TM 000 010

TABLE OF CONTENTS

I. Problems in Current Course Evaluation

II. Construction of the LAMMP Diagnostic Test

III. Analysis and Interpretation of Test Results

IV. Implications for the Construction of Content-Relevant Achievement Tests

Appendix

Bibliography

I. PROBLEMS IN CURRENT COURSE EVALUATION

In recent years there has been increased interest in the evaluation and improvement of instructional programs in an effort to improve education. Behavioral scientists evaluating instructional programs are coming to realize that the traditional "good" test, -- one which provides a large variance and distinguishes well among individuals -- is not necessarily appropriate for the evaluation and improvement of instruction. (See, for example, Cronbach, 1963; Guba, 1965; Husek, 1966; Skager, 1967; Stufflebeam, 1968.) The evaluation of individuals is very different from the evaluation of instruction, and a test which serves one function well does not necessarily serve the other equally well.

Standardized Achievement Tests in Evaluation Studies

There is a need for measuring instruments designed especially for the evaluation of instruction. The various types of educational situations calling for evaluation cannot always be met by one type of test. Most standardized achievement tests are designed to discriminate maximally among students at a given time. They measure the individual rather than the effectiveness of the course in producing desirable changes in the skills and understandings of the learners as a group. Whereas the standardized achievement test might give some indication of how much the student knows, it does not show which aspects of instruction are responsible for that knowledge. Yet when evaluating instruction, it is precisely

these aspects which are of vital interest. In the past evaluators have relied on the standardized achievement test to assess instructional programs. Unfortunately, these tests often fail to provide sufficient information for course improvement and are only of limited help in decision-making.

Total Scores versus Individual Items.

Total scores on standardized achievement tests tend to be highly stable over time and relatively insensitive to what occurs in the classroom. When instructional programs rather than individuals are being evaluated, however, the criterion of reliability no longer applies with the same force. The assignment of total scores, whatever their reliability, is of only limited help in improving the course.

It does not follow that standardized achievement tests have no place in evaluation studies, but the exclusive use of the total scores they provide is not very informative. Skager (1967) pointed out that the standardized achievement test, if analyzed at the item level, may actually be a useful tool in the evaluation of instruction. Whereas total scores may be affected very little by the program under study, the proportion passing individual items or clusters of items may change greatly from pretest to posttest. An analysis of individual items would provide information about the particular learning which has occurred, and this would be much more meaningful in evaluating instruction. Cronbach (1963) asserted that "outcomes of instruction are multi-dimensional...to agglomerate many types of post-course performance is a mistake, because

failure to achieve one objective is masked by success in another direction." Total scores on achievement tests are of little use when one wants to know exactly what has been taught. What is needed, then, in the evaluation of instruction, is some sort of diagnosis of instruction by analyzing student performance at the item level.

Content-Relevant Items

The items on a standardized achievement test are not necessarily relevant to an instructor's goals, and they seldom match the specific objectives of an experimental program. Husek (1966) feels that a course should be evaluated in terms of whether or not it meets its own objectives, and one of his criteria for the selection of items sensitive to changes in the learners is that they be related to the objectives of the course. Stake (1967) observed, moreover, that in the past there has been little attempt to measure the match between what an educator intends to do and what he actually does. He maintains that in evaluating instruction one is not so much asking whether results are reliable or valid but that what was intended did in fact occur.

Items should match instructional objectives without being limited by them; they should cover all important skills and understandings with which the students could reasonably be expected to be familiar.

Measurement of Change

Husek (1966) discusses new techniques of test construction which can meet the current needs of evaluators. He suggests an analysis of pretest-posttest differences in individual items to obtain specific

information about instructional effects. For this purpose, items should be selected which measure "changes in the students which have occurred during the course, not just final performance." These "change sensitive" items must be content-relevant, missed by most students at the beginning of the course, and passed by most students at its end. Should some of these items show little pretest-posttest change, it may not mean that they are insensitive to changes in student performance, but rather that the course inadequately covered the content or skills it was designed to measure.

Test Construction Procedures Needed in Evaluation

If tests are to be analyzed at the item level, items must be carefully chosen to ensure relevancy to instructional goals and broad coverage of important skills in the content area in question. A classification scheme which describes items in terms of what they measure would be a valuable guide in the selection of appropriate items and in matching instructional outcomes with the goals of the program and classroom activities.

This type of system would facilitate the interpretation of results, indicating whether specific course objectives have been met and answering precise questions about the adequacies and inadequacies of the course. If it is found, for example, that after appropriate periods of instruction the pupils are still weak in certain skills, the instructors would want to alter the program in such a way that these skills receive greater attention.

Evaluation of the Los Angeles Model Mathematics Project

The ensuing discussion suggests an alternative method of evaluating instructional programs which seems more appropriate than those traditionally employed. This method applies the preceding principles of course evaluation to the Los Angeles Model Mathematics Project (LAMMP), which is a compensatory program for disadvantaged junior high school pupils achieving one or more years below grade level in mathematics.

II. CONSTRUCTION OF THE LAMMP DIAGNOSTIC TEST

Collection and Classification of Items

The first step in the construction of a test to evaluate LAMMP was the collection of a very general pool of items representing nearly every possible mathematical subject matter for elementary and junior high school pupils. These items were organized into a meaningful system to classify mathematical content and the behaviors associated with that content, and to simplify the selection of items relevant to specific instructional goals. This system (Figure 1) has two dimensions, content (subject matter) and process (mental skills), and 96 cells. Each item in the item pool was classified into one of the cells according to the particular subject matter and skill it measures.

Though most items were quickly classified many seemed appropriate to more than one content area or mental process. This problem was temporarily solved by classifying such an item according to the process and content most essential to its solution. To a large extent these were subjective decisions and depended on the definitions given the content and process categories. Nevertheless, consistent application of the definitions following Figure 1 rendered the system sufficiently useful and reliable.

Selection of Items for the Test

The next problem was to select those items that would comprise the

TWO DIMENSIONAL CLASSIFICATION FOR ARITHMETIC ITEMS

Content

Integers--positive and negative whole numbers, zero.
Rational Numbers--fractions, decimals, percent.
Measurements--quantities (time, weight, distance, money, etc.), estimation.
Algebra--equations, functions, albegraic symbolism and terminology, cartesian coordinates.
Geometry--metric and non-metric geometry, plane, solid, and analytical geometry, proofs.
Numerals and Place Value--numeration systems, numerical bases, expanded notation, decimal notation, cardinal and ordinal numbers.
Number Theory--properties of numbers, inverse operations, algorithms, divisibility rules, equivalent & non-equivalent fractions, exponents, factors, even and odd numbers.
Sets--set notation, set theory, operations with sets, one to one correspondence, cartesian products.
Field Axioms and Principles--commutative, associative, distributative, and transitive principles, properties of equality, closure, identities, inverses, reciprocals.
Statistics--tables, graphs, averages, probability.
Word Problems--arithmetic problems as encountered in daily life situations.
Facts--definitions, terminology, formulas, symbols, basic arithmetic facts.

Process

Perceptual Skills--measurement and construction, visual recognition of concrete and figurative data, use of visual cues in problem solving.
Recognition and Recall--reading and recognizing numbers, memorization.
Computation--operations (addition, subtraction, multiplication, division), exponents and roots.
Conservation--preserving equality, conversions, translation.
Classification--groups and subsgroups, whole-part, association.
Seriation--ordering, series, sequences, pattern recognition, graduated order.
Relations--comparison, proportion, functions, correlations.
Application and Formal Logic--generalization, discovery, conclusions, evaluation, induction and deduction, analysis and synthesis, consistency, contradiction, negation, cause and effect.

Figure 1

LAMMP CLASSIFICATION MATRIX
FOR MATHEMATICS ITEMS

PROCESS	CONTENT: Integers	Rational Numbers	Measurement	Algebra	Geometry	Numerals & Place Value	Number Theory	Sets	Field Axioms & Principles	Statistics	Word Problems	Facts
Application	85	86	87	88	89	90	91	92	93	94	95	96
Relations	73	74	75	76	77	78	79	80	81	82	83	84
Seriation	61	62	63	64	65	66	67	68	69	70	71	72
Classification	49	50	51	52	53	54	55	56	57	58	59	60
Conservation	37	38	39	40	41	42	43	44	45	46	47	48
Computation	25	26	27	28	29	30	31	32	33	34	35	36
Recognition	13	14	15	16	17	18	19	20	21	22	23	24
Perceptual Skills	1	2	3	4	5	6	7	8	9	10	11	12

LAMMP Diagnostic Test. Most of the students in the target population came from poor, often broken homes and their experiences were very different from those of middle class children. An instrument was needed that would eliminate some of the extraneous variables usually associated with test performance. Vocabulary and sentence structure, instructions, and the recording of answers had to be simplified as much as possible, for many of the LAMMP pupils had serious reading disabilities.

Sociocultural factors also had to be considered. The content of each item had to be screened for subject matter with which the pupils may have had little experience. Math is sufficiently abstract; a frame of reference unfamiliar to the students would only make it that much more so. There is no claim that the LAMMP Diagnostic Test is culture free, but it represents an attempt, at least, to measure as much math and as little of everything else as possible.

Perhaps the most important factor to consider in selecting the items was the purpose for which the test was being prepared, namely, the evaluation of LAMMP in terms of its effects on the learners. The items had to be sensitive to these effects, that is, sensitive to change, and highly relevant to the content of the project if they were to provide valuable information when analyzing the test results.

With these criteria in mind, a pair of items was drawn from 40 carefully selected cells in the matrix representing most closely the instructional goals of the project or variables the evaluators most wished to observe. Each pair was randomly divided between test forms A and B (Appendix, pp. 34-45). Twelve simple computation problems from

cells 25 and 26 were added to each test form to provide some insight into the students' knowledge of basic skills.

Relevancy Ratings of the LAMMP Diagnostic Test Items

To assist in the analysis and interpretation of the observed effects of the project and to investigate the test's relevancy to the instructional goals of the project, a relevancy scale of the test's items was prepared.

These ratings were obtained shortly before the end of the project's first year. Experimental and comparison teachers were specifically asked to make a judgment on the extent to which instruction in their classes that year facilitated their students' ability to answer each of the LAMMP test's items correctly. (They were provided with the following scale:)

Instruction in my classes this year. . .

Definitely should not facilitate ability to answer	Probably would not facilitate	Uncertain	Probably should facilitate	Definitely should facilitate ability to answer
1	2	3	4	5

Administration of the LAMMP Diagnostic Test

The LAMMP Diagnostic Test was administered to experimental and comparison students in three junior high schools at the beginning of the fall term and again at the end of the spring term. Scores from one of the three project schools will be considered in this paper; they more

than adequately illustrate the methods used to analyze and interpret the results of the entire project.

The school's staff members randomly assigned incoming seventh grade students identified as eligible, that is, one or more years below grade level in mathematics achievement, into experimental and comparison classes. These classes were maintained intact throughout the year.

Comparison subjects were given traditional instruction in mathematics. The experimental groups were taught by novel methods, including the most recently developed instructional games, equipment, and machines.

III. ANALYSIS AND INTERPRETATION OF TEST RESULTS

Review of Total Scores

Mean pretest and posttest scores for Forms A and B of the LAMMP Diagnostic Test are given in Table 1:

Table 1

PRE AND POSTTEST PERFORMANCE FOR
EXPERIMENTAL AND COMPARISON SUBJECTS
ON LAMMP DIAGNOSTIC TEST (ITEMS 1-40)*

School	Pretest				Posttest			
	Experimental		Comparison		Experimental		Comparison	
	Fm A	Fm B	Fm A	Fm B	Fm A	Fm B	Fm A	Fm B
Belvedere								
mean	23.1	24.2	24.7	23.4	25.7	27.6	31.3	25.9
sigma	6.9	5.6	5.8	3.5	5.8	5.3	3.8	4.7
N	60	63	28	31	60	63	27	32

* The data reported in this paper are based only on those students who were in school for the entire school year.

There seem to be no consistent differences between experimental and comparison subjects with respect to overall gain. Both groups improved somewhat, and this is to be expected.

To assess LAMMP entirely on the basis of total scores would therefore be a highly discouraging prospect in that any successes of the

experimental program are obscured by the exclusive use of total scores. If the evaluator is to contribute to course improvement, he must provide the program developer with more useful information about the successes and weaknesses of the experimental program. As mentioned in the first section of this paper, a test analyzed at the level of the individual item or item cluster would help provide such information. This becomes especially useful when the item analysis, the pretest-posttest changes in individual items or item clusters, are compared to what is reported (teacher relevancy ratings of the test's items) and observed (classroom observations made by the evaluation team) to have occurred in the classroom.

The LAMMP Profiles of Pretest and Posttest Performance

When the items of a test are classified in a systematic way, as are those of the LAMMP Diagnostic Test, the data can be considered in a manner not usually possible with standardized achievement tests. The classification system in this study allows profiles of student performance to be drawn on the two dimensions of content and process from one short test. But when evaluating instruction, individuals need not be diagnosed; it is therefore not necessary to have a great number of items of any one type in order to insure adequate reliability.

The items of the LAMMP Test are classified on two dimensions, process and content. Profiles were therefore drawn for both dimensions providing a look at the same data from two points of view. They also provide a visual description of specific changes in student performance

over the year, information that total scores camouflage. A scrutiny of pretest-posttest differences in the height and shape of the profiles may lead to hypotheses concerning such questions as: Where are the experimental and comparison groups with respect to content and process before instruction? After instruction? Are there general trends noticeable from pretest to posttest? Is content differentially affected by instruction? Is process? Are the categories on a dimension related or are some very high and others low? Are the shapes of the profiles similar for experimental and comparison groups before instruction? After instruction? This way of examining data would seem to be highly appropriate to the evaluation and eventual improvement of instruction.

The LAMMP profiles (Figures 2-5) were prepared in the following manner: the percentage passing each item was determined for the experimental and comparison groups (Tables 2 and 3). The percentages were averaged over all the items at a given content or process category (column or row of the matrix). This was done for each group for the pretest and posttest separately. The items from both test Forms A and B at a given level were combined, thus doubling the number of items of any one type and increasing the reliability of the profiles. The N's indicate the number of items on which the mean percentage passing is based. Table 4 gives the cell numbers and the items within each cell. By referring to this table and to Figure 1 the test's items within each category may be identified.

Implications of the Teacher Ratings of Relevancy of LAMMP Diagnostic Test

The results of the teacher relevancy ratings (Table 5) indicate

Figure 2

PERCENT PASSING CONTENT CATEGORIES OF ITEMS ON THE LAMMP DIAGNOSTIC TEST, PRETEST

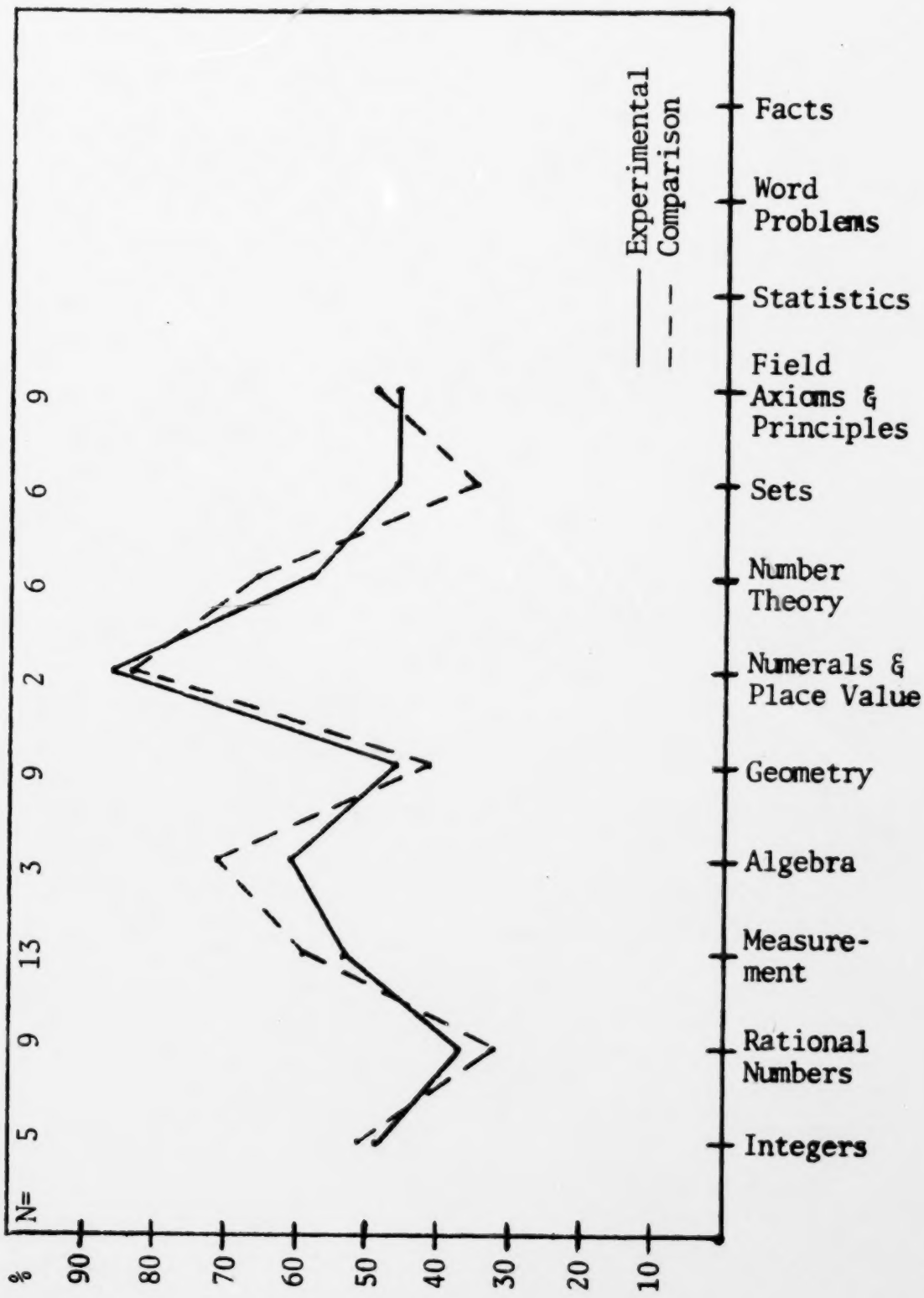


Figure 3

PERCENT PASSING CONTENT CATEGORIES OF ITEMS ON THE LAMMP DIAGNOSTIC TEST, POSTTEST

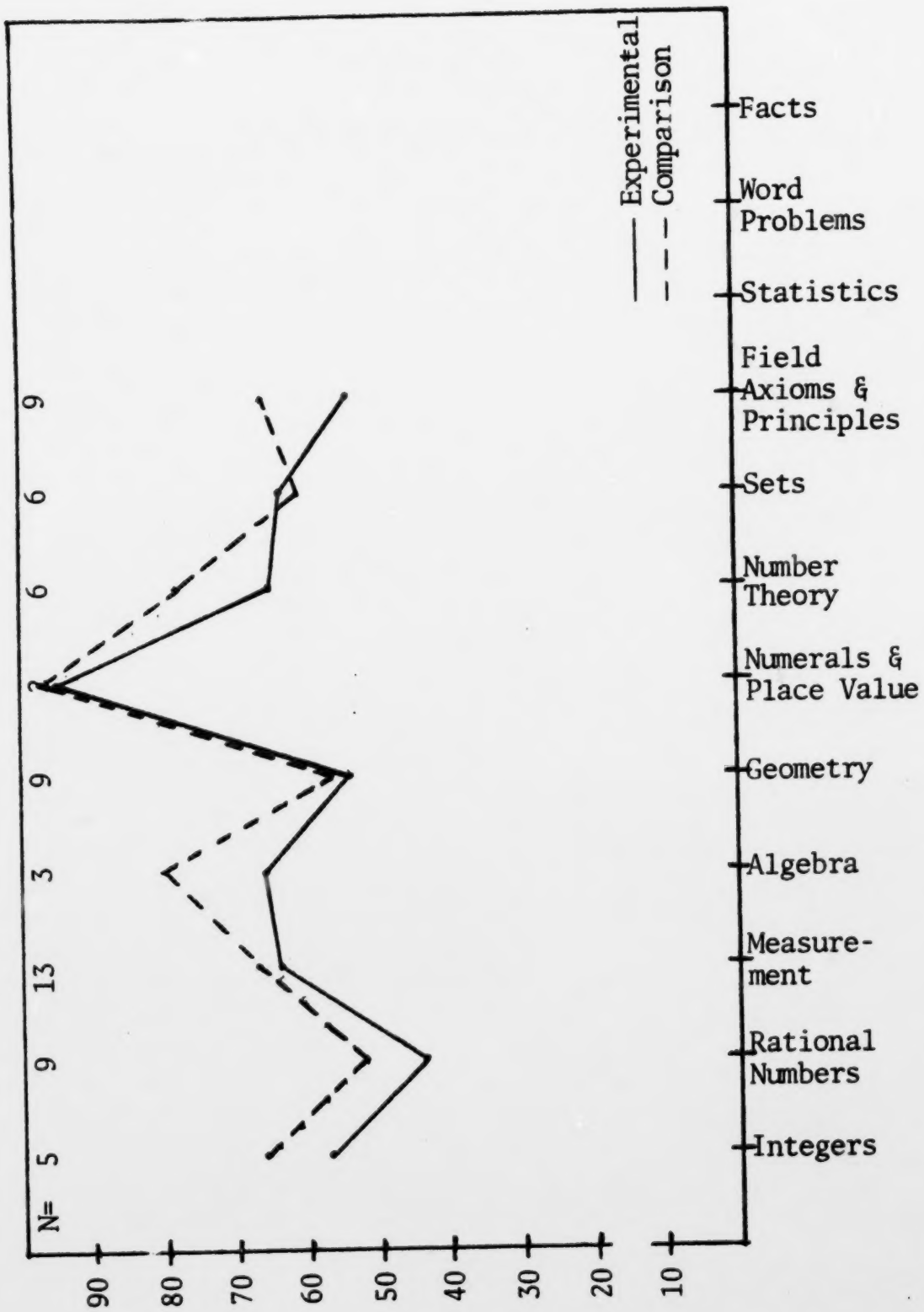


Figure 4

PERCENT PASSING PROCESS CATEGORIES OF ITEMS OF THE LAMP DIAGNOSTIC TEST, PRETEST

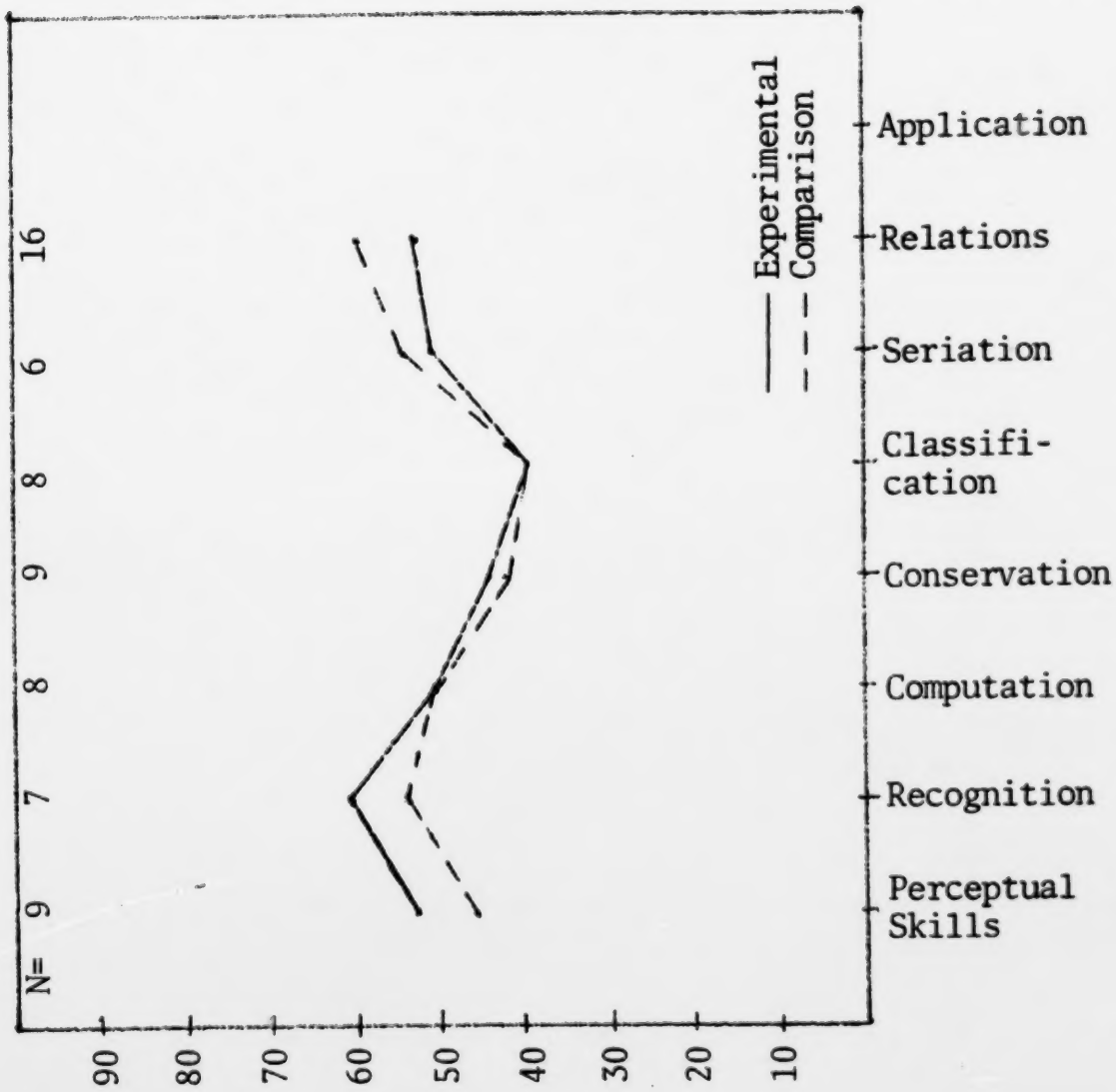


Figure 5

PERCENT PASSING PROCESS CATEGORIES OF ITEMS OF THE
LAMP DIAGNOSTIC TEST, POSTTEST

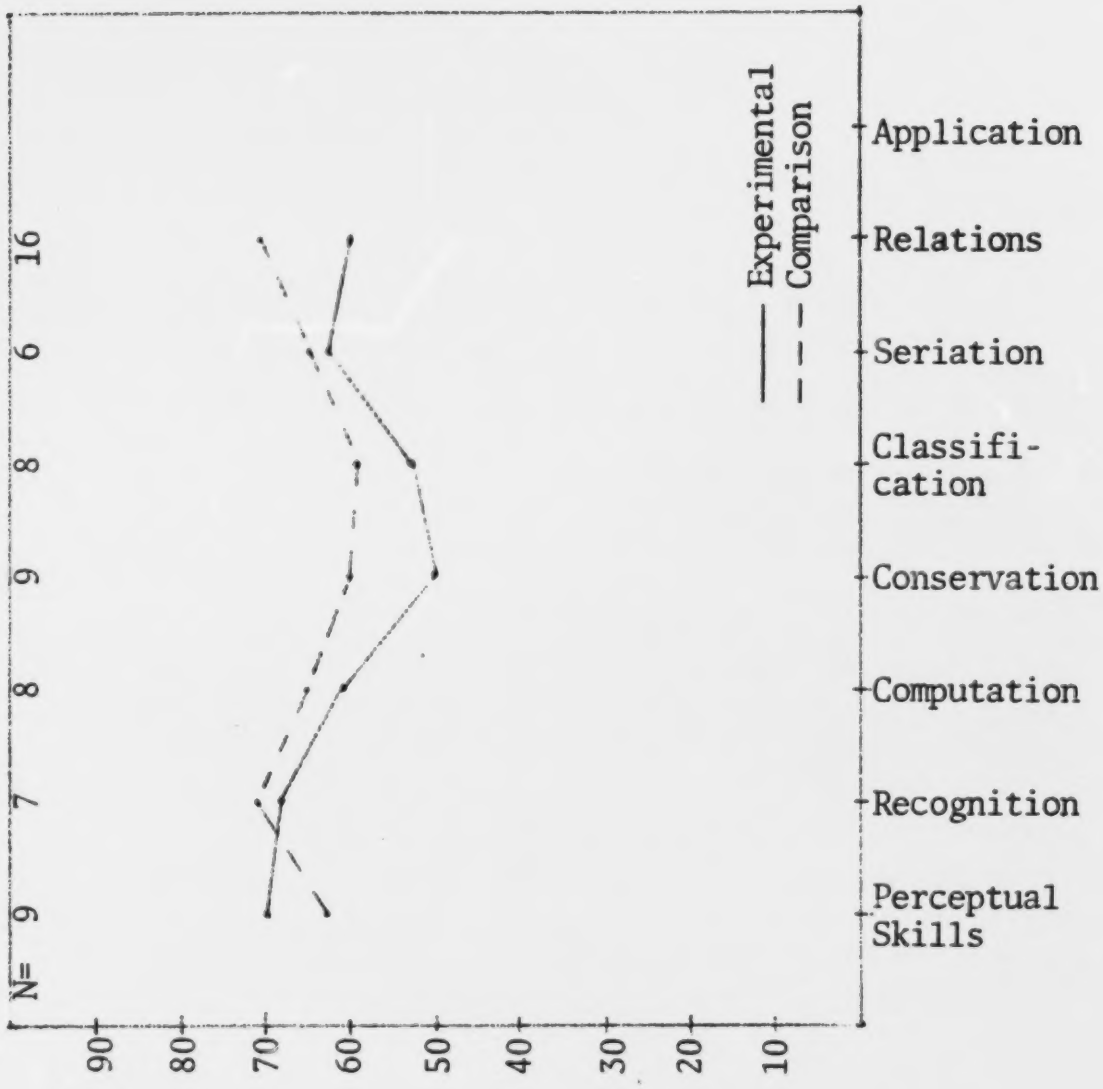


Table 2

PERCENT PASSING INDIVIDUAL ITEMS, PRETEST AND POSTTEST
(FORM A)

	EXPERIMENTAL			COMPARISON		
Items#	Pretest %correct	%difference	Posttest %correct	Pretest %correct	%difference	Posttest %correct
1	97	+3	100	100	0	100
2	42	+28	70	57	+10	67
3	87	+8	95	71	+25	96
4	28	-3	25	18	+23	41
5	52	+6	58	43	+42	85
6	35	+13	48	39	+2	41
7	25	0	25	21	+27	48
8	67	+6	73	71	+7	78
9	90	-3	87	96	-7	89
10	62	+8	70	79	+10	89
11	87	+6	93	100	0	100
12	32	-9	23	29	+15	44
13	88	+10	98	39	+61	100
14	55	+7	62	93	-4	89
15	35	+3	38	36	+5	41
16	83	+5	88	96	+4	100
17	20	+15	35	32	+1	33
18	57	-2	55	50	-43	93
19	57	+3	60	64	+17	81
20	78	+19	97	89	+11	100
21	52	+6	58	57	+32	89
22	78	+7	85	89	0	89
23	43	+3	47	50	+17	67
24	28	+14	42	29	+38	67
25	13	+10	23	11	+19	30
26	32	0	32	21	+53	74
27	98	-1	97	89	+11	100
28	57	+20	77	57	+32	89
29	73	+4	77	68	+25	93
30	60	+10	70	71	+25	96
31	50	+17	67	57	+32	89
32	48	+7	55	57	+17	74
33	70	+5	75	82	+11	93
34	25	0	25	21	+12	33
35	40	+3	43	64	+14	78
36	77	+8	85	82	+14	96
37	97	-4	93	89	+11	100
38	72	+18	90	82	+14	96
39	35	+5	40	21	+49	70
40	90	+2	92	100	-4	96

Table 3

PERCENT PASSING INDIVIDUAL ITEMS, PRETEST AND POSTTEST

(FORM B)

	EXPERIMENTAL			COMPARISON		
Items#	Pretest %correct	%difference	Posttest %correct	Pretest %correct	%difference	Posttest %correct
1	75	+3	78	52	+26	78
2	48	+22	70	42	+17	59
3	70	+14	84	68	+4	72
4	14	+29	43	10	-1	9
5	62	+25	87	48	+17	75
6	95	-1	94	77	+17	94
7	32	+11	43	32	+15	47
8	8	+17	25	13	+6	19
9	94	-8	86	84	+7	91
10	94	+6	100	100	0	100
11	83	-4	87	87	-9	78
12	54	+29	83	39	+36	75
13	97	+3	100	100	0	100
14	60	+24	84	55	+10	66
15	68	+22	90	77	+14	91
16	59	+12	71	48	+27	75
17	71	+4	75	61	+2	63
18	35	-6	29	16	+31	47
19	35	+13	48	32	+6	38
20	76	+5	81	71	+13	84
21	94	0	94	94	-10	84
22	62	+14	76	87	+1	88
23	98	+2	100	94	+3	97
24	98	0	98	100	-6	94
25	59	0	59	65	-2	63
26	37	+15	52	13	+21	34
27	51	+19	70	77	-24	53
28	86	+9	95	94	-6	88
29	17	+5	22	3	+16	19
30	86	-5	81	77	+10	88
31	10	+23	33	13	+6	19
32	73	+11	84	97	-13	84
33	11	+3	14	6	0	6
34	41	+10	51	48	-7	41
35	53	+3	57	35	+9	44
36	81	-6	75	84	+4	88
37	95	-1	94	100	0	100
38	67	+4	71	68	-2	66
39	37	+11	48	35	+6	41
40	32	+1	33	35	+3	38

Table 4. Cell Numbers and Corresponding Items

Cell#	Item#	
	A	B
2	1	1
2	40	37
3	2	2
5	3	3
5	4	4
8		12
8		14
13	6	7
15	8	9
18	11	6
18	20	
19	12	
24	13	13
27	9	15
27	14	
29	15	16
31	16	17
32		5
33	17	18
38	7	8
38	18	19
39	19	20
39	34	35

Cell#	Item#	
	A	B
42		11
42		21
45	30	28
49	23	24
53	24	25
56	5	26
56	25	
57	26	27
61	27	23
62	28	29
63	22	10
63		31
69	29	32
73	31	30
74	32	33
75	33	34
76	35	36
76	36	38
77	37	40
79	10	
79	21	22
79	38	
81	39	39

that the teachers found the test as a whole to be relevant to what they taught that year. It is therefore legitimate to use the ratings as a tool in the interpretation of instructional outcomes.

Table 5

MEAN RATINGS OF RELEVANCY
OF LAMMP DIAGNOSTIC TEST ITEMS*

School	Experimental		Comparison	
	Fm A	Fm B	Fm A	Fm B
Belvedere	3.71 (5)	3.81 (5)	4.28 (3)	4.23 (3)

Considering all the profiles, it is interesting that for the most part, and especially for the experimental group, process and content areas which are low on the pretest are also low on the posttest. Does this reflect the fact that teachers failed to correct student weaknesses and to a certain extent emphasized what students already knew? And does it mean that experimental teachers did this to a greater extent than did comparison teachers? There is evidence to support these assumptions. Table 6 gives the correlation coefficients between the proportion of

---

* The number in parenthesis below the means indicates the number of ratings on which each was based.

students answering the items correctly on the pretest and the teachers' ratings of relevancy. Note that all the coefficients are positive and that they are consistently higher for the experimental teachers. The teachers may thus have indicated that they taught for student strengths rather than student weaknesses; and if this were the case, it would be unrealistic to expect significant improvement in student performance, for instruction would have emphasized what the students already knew. Perhaps students in general are taught what they know. Perhaps this is what instructors find easiest to teach. On the other hand, perhaps the LAMMP teachers rated not what occurred in the classroom but their own perceptions of student strengths in an unconscious attempt to take credit for those areas in which their students would do well. This would explain the higher correlations on the part of experimental teachers, for they might understandably feel threatened by the attention given the achievement of their students. This problem deserves further study.

Table 6

RELATIONSHIPS BETWEEN PROPORTION ANSWERING ITEMS OF LAMMP DIAGNOSTIC TEST CORRECTLY AT PRETEST AND TEACHER'S RATINGS OF RELEVANCY TO INSTRUCTION

School	Experimental		Comparison	
	FM A	Fm B	Fm A	Fm B
Belevedere	.25	.62	.14	.30

Analysis of Item Clusters Represented in the Profiles

The experimental and comparison group profiles are very similar on the pretest; they are close together and quite parallel. The two groups are comparable with respect to specific strengths and weaknesses as well as total scores. Further study of the profile reveals that in every category both groups improved (Figures 2-5).

On the posttest, experimental and comparison content profiles appear less equivalent. Although differences remain small, the comparison group falls below the experimental group only once. Moreover, the comparison group shows relatively greater improvement in those areas where both groups are low on the pretest. For the four lowest categories on the pretest (Rational Number, Geometry, Sets, Field Axioms and Principles) the experimental group improved an average of 10%, compared to 20% for the comparison group. The corresponding figures for the four highest levels on the pretest (Algebra, Numerals and Place Value, Number Theory, Measurement) are 8.5% and 11%; the ratio is much smaller. The relevancy ratings help explain this phenomenon; as mentioned above, experimental teachers indicated a greater neglect of student weaknesses in their ratings than did comparison teachers. Yet, in spite of the positive correlation between relevancy and student strengths, both groups show greater change where they began low than where they began high. This must be expected, for the test's ceiling prevents students from improving as much as they might have had more difficult items been selected; and the effect of regression is to draw extreme scores back towards the mean.

Considering the content categories individually, one finds a much greater increase for the comparison group on Rational Numbers than the experimental group (20% increase compared to 7% increase). Classroom observers noted some experimental teachers working with fractions; decimals, however, seemed to be neglected. The few comparison teachers observed appeared to place a greater emphasis on all types of rational numbers. The ratings of relevancy support the observations and the profile data. Experimental teachers gave these items an average of 3.1 points while the comparison teachers' average rating was 4.4 points on a scale of 5.

In view of the high pretest scores on Numerals and Place Value, both groups show a surprising gain, especially the comparison group. As expected, these items were rated as very relevant by both groups of teachers, and the comparison teachers' average rating was the higher of the two (5.0 and 4.3).

As with the content profiles, the process profiles overlap and are quite parallel at pretest, while on the posttest the comparison group moves ahead of the experimental group. Of course, since process cuts across content, overall change must be in the same direction on both dimensions.

In recognition the experimental group improved only 7% compared to the comparison group's 18%. Classroom observation checklists indicate that the experimental teachers emphasized instructional games, student interaction, spirited activities, and discovery methods. With this came a corresponding de-emphasis of drill and rote recall. The few

observations available for comparison classes suggest that these teachers did, on the other hand, demand considerable memorization from their pupils. The relevancy ratings again support the observed results; every recognition item was rated as more relevant by comparison teachers than by experimental teachers. The averages are 4.8 and 4.1, correspondingly.

In conservation, as in recognition, the comparison group began below the experimental group and then surpassed it. Observations do not indicate a great deal of work with equalities and conversions (such as decimals to fractions or percents) in experimental classes. The relevancy ratings concur. The mean rating for conservation items is 2.9 for experimental teachers and 4.5 for comparison teachers. It is hardly surprising then that comparison subjects show an 18% gain compared to the experimental subjects' 6% gain.

Concluding Comments and Recommendations

The preceding pages illustrate how many positive aspects of an experimental program may be discovered by analyzing pretest-posttest changes at the item level and comparing the observed changes to reported and observed classroom activities. These aspects, worth preserving, are obscured by total scores, as are the negative effects of a program, which can be corrected only if they are known.

When the evaluator determines the specific advantages and inadequacies of an instruction program and communicates these to the program developers, he is doing more than passing judgment; he is giving counsel and advice in the form of diagnostic information and thus contributing

to course improvement.

Evaluation can further contribute to course improvement by providing teachers and program developers with feedback on student pretest performance at the beginning of the course. They could be given a matrix such as that developed for LAMMP and the corresponding pretest content and process profiles. This would aid in the selection of instructional goals, which need not be the same for every class, and which would probably differ significantly from school to school. The selected objectives could then be tested repeatedly throughout the school year. The results would guide instructional emphases in light of student progress. With respect to the progressive improvement of an experimental program, this seems to be an excellent plan. At the same time, the final posttest would clearly reveal those goals which have or have not been attained; the total scores would give an indication of the overall effectiveness of the program.

It is also suggested that in future analyses, profiles of teacher relevancy ratings be prepared on the same basis that the profiles of student performance were prepared. This would be of practical value in facilitating the interpretation of outcomes. The rating profiles could also be of interest to one wishing to study the relationships between what a teacher says he does, what he actually does, and the effects on student performance.

IV. IMPLICATIONS OF THE LAMMP CLASSIFICATION SYSTEM

Diagnosis of Instruction

Each cell in the classification matrix represents a type of mental process necessary to solve certain mathematical problems and a content area to which this process is applied. The cells describe items and/or a class of behavioral objectives. Selection of content relevant items to measure specific instructional objectives is thereby greatly facilitated. For this purpose one would select items from those cells matching the objectives.

Should one wish the test to serve a primarily diagnostic function, every cell in the matrix should be sampled. As a pretest this would serve to locate student weaknesses which would subsequently be emphasized in the classroom. These areas could be measured repeatedly throughout the course, the outcome guiding instruction, until the students have reasonably mastered the material. As a posttest, one would be assured of measuring all or almost all of the important instructional outcomes in the particular subject matter.

By organizing instructional objectives, the common tendency to overemphasize certain topics or behaviors (i.e., recall) to the detriment of others is likewise avoided. Of course, should the test builder wish to emphasize particular areas, even to the exclusion of others, this, too, is made easier. The advantage of selecting every item for a known

purpose is that the final results may be interpreted in terms of the preceding learning activities.

Importance of Measuring Process

The greater ease with which an item may be classified in terms of content probably explains why many tests thoroughly cover the content area in question. But constructing a test solely with respect to contact does not ensure an adequate coverage of important skills and processes. It is, perhaps, for this reason that the items on many mathematics tests, including standardized achievement tests, are at the computation or recognition level. Items may "look" very different because they measure different content, when actually the mental process involved is the same. A classification scheme based on process as well as content would likely lead to a better balance between the two, not only in measurement, but in the curriculum as well.

Perhaps many pupils are underachieving not merely because they have not learned the content of their courses but because they are weak in certain skills and abilities prerequisite to achievement in those content areas. This would make it essential not to overlook the diagnosis and instruction of processes and skills. And when these are known, specified, and categorized, as is content, this task is greatly facilitated.

Classifications for Other Subject Matters

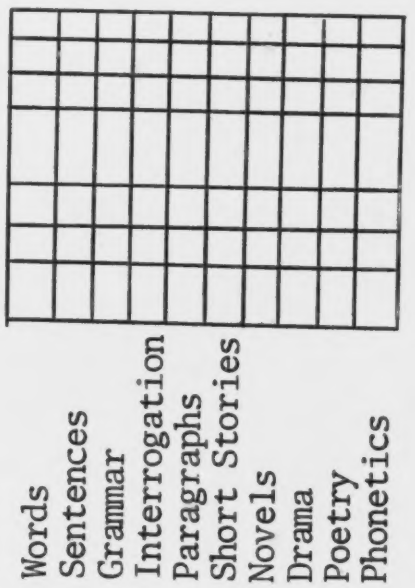
Though the LAMMP matrix was constructed specifically for mathematical subject matter and skills, the idea behind the system has more

general application. One might argue that math more readily lends itself to a detailed analysis of the behaviors involved, but this does not preclude the system's usefulness in other fields.

A classification in other subjects would function as it does in mathematics, providing many of the same services. In foreign languages, for example, the schema might look roughly something like this:

Process

Translation
Writing
Reading
Speaking and
Pronunciation
Listening Comprehension
Repetition
Imitation



Contribution to Basic Research

A possible consequence of the item classification strategy has to do with research in cognitive structures and levels of cognitive development. The researcher would likely find a breakdown of behaviors in a content area of significant value. Of course, there is no claim that the process levels of the LAMMP matrix represent, at this time, anything more than a very useful way of describing skills used to solve mathematical test items. On the other hand, additional studies and research

might find these categories to be related to theories about cognitive structure. The profiles would then have a purpose beyond course improvement and student assessment. For example, the researcher might want to look for interactions between profiles in one subject matter and similar profiles in other subject matters. He might be interested in comparing profiles of cognitive structure with characteristics such as general intelligence, special talents, or the ability to profit from various instructional techniques and materials. He may wish to know whether profiles of cognitive development fluctuate with age or socioeconomic class, whether the actual shape of the profile or merely its height varies with intelligence, and at what age, if at all, one's cognitive structure becomes fixed. He may wonder what the consequences of instruction might be if it were geared to process rather than content.

Many Ways to Represent Cognitive Structure

Many systems have been developed for the classification of mental processes (e.g., Bloom, 1955; Bruner, 1956; Gagné, 1959; Guilford, 1959). None of them are ideally suited to all situations. Whereas Guilford's model may be more valuable to the researcher, the classroom teacher would probably consider Bloom's Taxonomy more useful.

Unlike the LAMMP classification, most of these systems are independent of subject matter. That is, their operations are generally applicable to all fields of learning. Dressel (1949) devised a two-dimensional plan for the selection of test items which does take subject matter into account; but though he places content on one dimension, on the other

are only very general instructional objectives, such as "to achieve a knowledge and understanding of biological concept" and "ability to read historical materials." These are very general activities involving the content area itself. The LAMMP process dimension represents an attempt to describe items in terms of more fundamental operations; and its cells, rather than the dimension, represent the instructional objectives.

Conclusion

Whatever the system, it must ultimately be judged in terms of the purpose for which it was developed. The LAMMP classification was devised to help in the construction and analysis of content-relevant mathematics achievement items for the evaluation of an instructional program; this it did. It was constructed in the hope that it may have even broader applications; it may.

If a system is to be psychologically meaningful, it must be verified empirically. For this reason the LAMMP classification must remain tentative with respect to basic research in cognitive development. Its present value can only be in terms of its rationale rather than the intrinsic worth of its categories.

The method described in this paper for the construction, analysis, and interpretation of a test to evaluate instructional programs represents a different and alternative approach to the traditional over-reliance on standardized achievement tests and the total scores they provide. This method led to a content-relevant, change-sensitive test which was applied to the evaluation of the Los Angeles Model Mathematics

Project. The pretest and posttest results were analyzed at the item level and interpreted in terms of specific classroom activities. This type of approach may contribute to the effectiveness of course evaluation and be a step in the direction of providing services that today are greatly needed.

Appendix

THE LAMMP DIAGNOSTIC TEST
(Forms A and B)

FORM A

NAME _____

SCHOOL _____

LOS ANGELES MODEL MATHEMATICS DIAGNOSTIC TEST

CIRCLE THE CORRECT ANSWER

1. Which drawing is divided into thirds?

a.

b.

c.

d.

2. Without measuring, tell how long you think this line is.

a. 1 inch
b. 2 inches
c. 3 inches
d. 4 inches

3. Which figure does not belong?

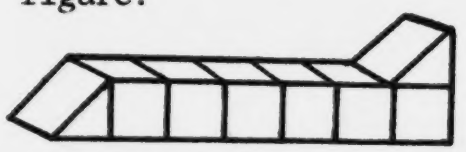
a.

b.

c.

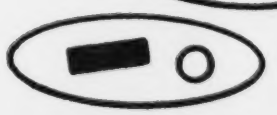
d.

4. What is the number of cubic units (volume) of the figure?

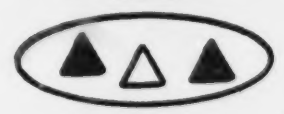


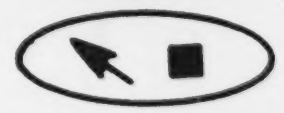
a. 5
b. 6
c. 7
d. 8

5. Which of the sets a, b, c, or d is a subset of the following set?

a. 

b.

c. 

d. 

6. Circle the numeral which represents sixteen hundred fifty-nine.

a. 1,659
b. 160,059
c. 16059
d. 1,600,059

7. Which numeral is equal to .05?

   a. .005
   b. .050
   c. .500
   d. .055

8. How many ounces in a pound?

   a. 8
   b. 10
   c. 16
   d. 32

9. How many eggs are there in one-half dozen?

   a. 3
   b. 24
   c. 12
   d. 6

10. Which numeral goes in the box?

$$\frac{17}{45} - \square = 0$$

   a. $\frac{17}{45}$

   b. $\frac{45}{17}$

   c. 0

   d. 1

11. In the numeral 5,271 the 5 represents:

   a. 5 thousands
   b. 5 hundreds
   c. 5 tens
   d. 5 ones

12. Circle the <u>odd</u> numeral:

   a. 70
   b. 683
   c. 516
   d. 3152

13. When you see 13 5 you are to:

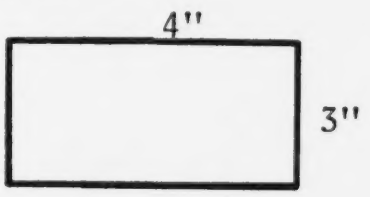
   a. Multiply
   b. Substract
   c. Add
   d. Divide

14. Multiply:

   2 feet 3 inches
   X 3

   a. 6 feet 6 inches
   b. 4 feet 9 inches
   c. 5 feet 6 inches
   d. 6 feet 9 inches

15. What is the perimeter (distance around) of this rectangle?



   a. 7 inches
   b. 12 inches
   c. 14 inches
   d. 25 inches

16. Which number goes in the box?

$$8 \times \square = 32$$

   a. 3
   b. 2
   c. 4
   d. 6

17. Substract:

   9 - (6 - 3) =

   a. 0
   b. 3
   c. 6
   d. 9

18. Which decimal numeral is equal to 26%?

    a. .26
    b. .52
    c. .62
    d. 2.6

19. Which statement is true?

    a. 120 minutes = 3 hours
    b. 120 minutes = 2 hours
    c. 120 minutes = $1\frac{1}{2}$ hours
    d. 120 minutes = 4 hours

20. In the numeral 42, 355 the <u>3</u> represents:

    a. 3 thousands
    b. 3 hundreds
    c. 3 tens
    d. 3 ones

21. Which numeral goes in the blank?

    649.88 X _____ = 0

    a. 1
    b. 649.88
    c. -649.88
    d. 0

22. If May 4th is on a Monday, then May 8th is on a:

    a. Wednesday
    b. Thursday
    c. Friday
    d. Saturday

23. Which set has only numbers between ten and fifteen?

    a. {7, 8, 9, 10}
    b. {11, 12, 16, 18}
    c. {1, 5, 10, 15}
    d. {11, 12, 13, 14}

24. Squares, rectangles, and parallelograms are alike in that:

    a. all have four right angles
    b. all have four sides of equal length
    c. all have four sides
    d. all are three-dimensional figures

25. Set R = {a, c, e, i, m, o} Which set is a <u>subset</u> of Set R?

    a. {a, b, c, d, e}
    b. {i}
    c. {a, e, i, o, u}
    d. {a, e, f}

26. Which numeral goes in the blank?

    .375 X (.5 X .16) = (.375 X .5) X____

    a. .16
    b. .375
    c. .5
    d. .380

27. Which numeral is missing?

    8, 7, __, 5, 4

    a. 8
    b. 6
    c. 4
    d. 3

28. Which numeral has been left out?

    1, $\frac{1}{2}$, $\frac{1}{4}$, __, $\frac{1}{16}$

    a. $\frac{1}{6}$
    b. $\frac{1}{10}$
    c. $\frac{1}{8}$
    d. $\frac{1}{12}$

29. Which numeral goes in the blank?

30 X 15 = 15 X _____

a. 15
b. 45
c. 30
d. 60

30. Which numeral goes in the blank?

271 X ____ = 271

a. 1
b. 0
c. 10
d. $\frac{1}{2}$

31. Which statement is true?

a. 9 - 4 is equal to 6
b. 9 - 4 is less than 6
c. 9 - 4 is greater than 5
d. 9 - 4 is not equal to 5

32. Which statement is true?

a. $\frac{1}{4}$ is greater than $\frac{1}{2}$

b. $\frac{1}{4}$ is equal to $\frac{1}{2}$

c. $\frac{1}{4}$ is less than $\frac{1}{2}$

d. $\frac{1}{4}$ is equivalent to $\frac{1}{2}$

33. What is the smallest number you can get when throwing TWO dice?

a. 1
b. 2
c. 3
d. 4

34. 3 inches is the same as:

a. $\frac{1}{6}$ foot

b. $\frac{1}{4}$ foot

c. $\frac{1}{3}$ foot

d. $\frac{1}{2}$ foot

35. If 8X = 40, then X must equal

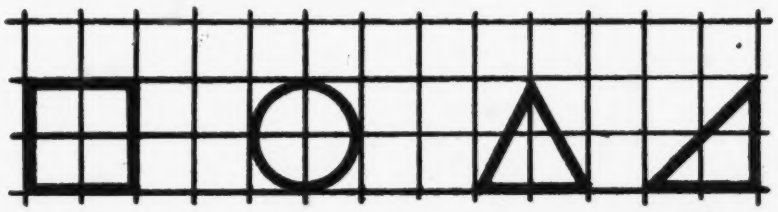
a. 320
b. 5
c. $\frac{1}{5}$
d. 5X

36. Which numeral goes in the box?

$\frac{3}{5} - \frac{\square}{5} = \frac{1}{5}$

a. 1
b. 2
c. 4
d. 5

37. Which figure has the greatest area?



a. b. c. d.

38. Which numeral goes in the blank?

.49 - _____ = 0

a. 1
b. 0
c. .51
d. .49

39. Which sign goes in the circle?

350 - 50 ◯ 196

a.
b. -
c.
d. =

40. Find $\frac{27}{39}$ in the list below

a. $\frac{29}{37}$

b. $\frac{37}{29}$

c. $\frac{39}{27}$

d. $\frac{27}{39}$

DO THESE PROBLEMS:

$$\begin{array}{r} 6 \\ +\ 12 \\ \hline \end{array}$$

$$\frac{1}{3} + \frac{1}{3}$$

$$\begin{array}{r} 8.9 \\ +\ 2.0 \\ \hline \end{array}$$

$$\begin{array}{r} 6 \\ \times\ 9 \\ \hline \end{array}$$

$$\frac{1}{2} \times \frac{3}{4} =$$

$$\begin{array}{r} 1.9 \\ \times\ .3 \\ \hline \end{array}$$

$$\begin{array}{r} 13 \\ -\ 7 \\ \hline \end{array}$$

$$\begin{array}{r} 2\frac{1}{2} \\ -\ 1\frac{1}{4} \\ \hline \end{array}$$

$$\begin{array}{r} 4.8 \\ -\ .6 \\ \hline \end{array}$$

$$5\overline{)35}$$

$$\frac{3}{4} \div \frac{2}{3} =$$

$$1.4\overline{)1.20}$$

FORM B

NAME _____

SCHOOL _____

LOS ANGELES MODEL MATHEMATICS DIAGNOSTIC TEST

CIRCLE THE CORRECT ANSWER

1. How much of this picture is dark?

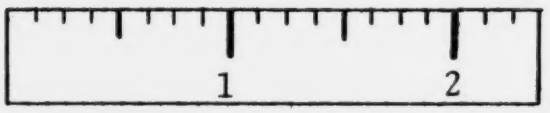
a. $\frac{3}{4}$

b. $\frac{1}{4}$

c. $\frac{4}{3}$

d. $\frac{4}{4}$

2. How long is the line shown above the ruler?



a. $1\ \frac{3}{8}$ in.

b. $1\ \frac{1}{2}$ in.

c. $1\ \frac{5}{8}$ in.

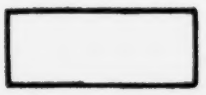
d. $1\ \frac{5}{6}$ in.

3. Which figure does not belong?

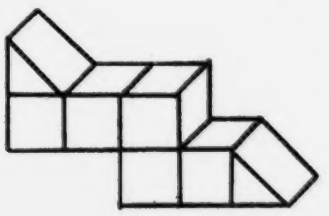
a.

b.

c.

d. 

4. What is the number of cubic units (volume) of the figure?



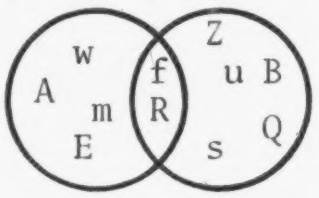
a. 4

b. 5

c. 6

d. 7

5. What is the intersection of these two sets?



a. A, m, E, w, s, Z
b. f, R
c. f, R, z, s, Q, B, u
d. A, E, R, S, U

6. In the numeral 1,720 the 7 represents:

a. 7 thousands
b. 7 hundreds
c. 7 tens
d. 7 ones

7. Which numeral represents twelve hundred ten?

a. 1,210
b. 12,010
c. 120,010
d. 1,200,010

8. Which numeral represents .4?

a. .004
b. .040
c. .400
d. .41

9. How many inches in a foot?

a. 1
b. 5
c. 12
d. 24

10. In which space should Thursday be?

Sun	Mon	a	b	c	d	Sat

a. a
b. b
c. c
d. d

11. The numeral 57 represents:

a. 5 + 7
b. 5 X 10 + 7
c. 5 X 7
d. 5 X 70

12. Which sets are equal?

W = {□ △}

X = {△ ○}

Y = {○ □}

Z = {△ □}

a. Z and Y
b. W and Y
c. W and Z
d. X and Z

13. When you see 18 X 6 you are to:

a. add
b. subtract
c. multiply
d. divide

14. Which sets have the same number of members?

W = {□ □ □}

X = {○ ○ ○ ○}

Y = {△ △ △}

Z = {▭}

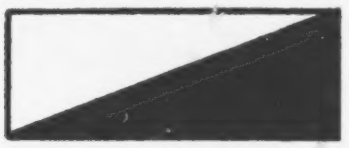
a. W and Y
b. W and Z
c. Y and Z
d. W and X

15. Multiply:

4 pounds 2 ounces
X 2

a. 12 pounds
b. 8 pounds 2 ounces
c. 4 pounds 4 ounces
d. 8 pounds 4 ounces

16. The area of this rectangle is 6 square inches. What is the area of the dark part?



a. 1 sq. in.
b. 2 sq. in.
c. 3 sq. in
d. 4 sq. in.

17. Which numeral goes in the box?

□
+ .3
2.4

a. 1.1
b. .1
c. 2.7
d. 2.1

18. Substract:

15 - (5 - 3) =

a. 8
b. 7
c. 12
d. 13

19. Which numeral represents 4 7 ?
10

a. 47
b. 4.7
c. 4.07
d. 470

20. Which statement is true?

a. 12 inches = 2 feet
b. 20 inches = 2 feet
c. 24 inches = 2 feet
d. 48 inches = 2 feet

21. The numeral 29 represents:

a. 2 + 9
b. 2 X 9
c. 2 X 10 + 9
d. 2 ÷ 9

22. Which numeral goes in the blank?

3568 X _____ = 0

a. 1
b. 0
c. 3568
d. -3568

23. Which numeral has been left out?

5 10 __ 20 25 30

a. 11
b. 15
c. 19
d. 100

24. Which set has only numbers that are smaller than 7?

a. {7, 8, 9, 10}
b. {3, 5, 8}
c. {3, 4, 5}
d. {1, 6, 9, 10}

25. Which one of these can sometimes be a square?

a. circle
b. triangle
c. line
d. rectangle

26. Which statement is true?

P = {4, 7, 28, 96}
Q = {7, 96}

a. P and Q are equal sets
b. P is a subset of Q
c. Q and P are equivalent sets
d. Q is a subset of P

27. Which numeral goes in the blank?

359 X (6000 X 5) = (359 X __) X 5

a. 6000
b. 30,000
c. 1795
d. 21,540

28. Which numeral goes in the blank?

765 + _____ = 765

a. 1
b. 765
c. -765
d. 0

29. Which group of fractions are in order from least to greatest?

a. $\frac{4}{4}, \frac{2}{4}, \frac{3}{4}, \frac{1}{4}$

b. $\frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{6}$

c. $\frac{2}{2}, \frac{3}{3}, \frac{4}{4}, \frac{5}{5}$

d. $\frac{1}{4}, \frac{1}{3}, \frac{1}{2}, \frac{3}{4}$

30. Which statement is true?

a. 7 + 3 is equal to 12
b. 7 + 3 is less than 12
c. 7 + 3 is greater than 10
d. 7 + 3 is not equal to 10

31. If you have 17¢ in coins, what is the smallest number of coins you can have?

a. 2
b. 3
c. 4
d. 8

32. Which numeral goes in the blank?

$$_____ + 75 = 75 + 16$$

a. 16
b. 75
c. 91
d. 59

33. Which numeral represents the largest number?

a. .605
b. .65
c. .617
d. .0607

34. One yard is:

a. shorter than 2 feet
b. equal to 2 feet
c. shorter than 4 feet
d. longer than 4 feet

35. 15 minutes is the same as:

a. $\frac{1}{6}$ hour
b. $\frac{1}{4}$ hour
c. $\frac{1}{2}$ hour
d. $\frac{1}{3}$ hour

35. If 5 + X = 8, then X must equal:

a. 3
b. 13
c. 31
d. 13 X

37. Find 2.75 in the list below:

a. 27.5
b. 2.75
c. .275
d. 275

38. Which numeral goes in both boxes?

$$\frac{4}{\square} - \frac{1}{\square} = \frac{3}{5}$$

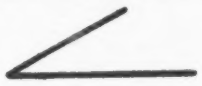
a. 1
b. 3
c. 4
d. 5

39. Which sign goes in the circle?

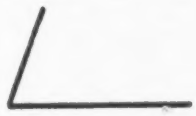
4 X 7 ◯ 3 X 8

a. >
b. =
c. +
d. <

40. Which angle has the largest measure?

a. 

b.

c. 

d.

DO THESE PROBLEMS:

$$\begin{array}{r} 13 \\ +\ 14 \\ \hline \end{array} \qquad \begin{array}{r} 3.01 \\ +\ 1.02 \\ \hline \end{array} \qquad \frac{1}{4} + \frac{2}{4} =$$

$$\begin{array}{r} 11 \\ -\ 6 \\ \hline \end{array} \qquad \begin{array}{r} 5.7 \\ -\ .2 \\ \hline \end{array} \qquad 3\ \frac{2}{3} - 2\ \frac{1}{3} =$$

$$\begin{array}{r} 8 \\ \times\ 7 \\ \hline \end{array} \qquad \begin{array}{r} 2.3 \\ \times\ 2.0 \\ \hline \end{array} \qquad \frac{2}{3} \times \frac{1}{2} =$$

$$7\overline{)70} \qquad 1.5\overline{)2.50} \qquad \frac{1}{2} \div \frac{2}{3} =$$

BIBLIOGRAPHY

Bloom, B.S. (Ed.) Taxonomy of educational objectives; Handbook I: Cognitive domain. New York: David McKay, 1956.

Bruner, J.S., Goodnow, J.J., & Austin, G.A. A study of thinking. New York: Wiley, 1956.

Cronbach, Lee J. Course improvement through evaluation. Teacher's College Record, 1963, 64, 672-683.

Dressel, P.L., et al. Comprehensive examinations in a program of general education. East Lansing: Michigan State University Press, 1949.

Epstein, M.G. Computer assembly of tests. Educational Testing Services, Test Development Division, Princeton, New Jersey. (Address delivered to the Military Testing Association, Toronto, Canada, September 27, 1967.

Flavell, J.H. The developmental psychology of Jean Piaget. Princeton, N.J.: Van Nostrand, 1964.

Gagné, R.M. Problem solving and thinking. Annual Review of Psychology 1959, 10, 147-172. Palo Alto: Annual Reviews, Inc.

Guba, E.G. Evaluation and the process of change. Notes and Working Papers Concerning the Administration of Programs authorized under Title III of Public Law 89-10, the Elementary and Secondary Education Act of 1965 as amended by Public Law 89-750, April 1967, p. 312.

Guilford, J.P. Three faces of intellect. American Psychologist, 1959, 14, 469-479.

Husek, T.R. Different kinds of evaluation and their implications for test development. (Paper read before the Fiftieth Annual Meeting of the American Educational Research Association, Chicago, Feb. 19, 1966.)

Isaac, N. The growth of understanding in the young child: A brief introduction to Piaget's work. London: Educational Supply Association, 1963.

_____. New light on children's ideas of number: The work of Professor Piaget. London: Educational Supply Association, 1960.

Scriven, M. The Methodology of Evaluation. AERA Monograph Series on Curriculum Evaluation, 1967, I, 39-89. Chicago: Rand McNally.

Skager, R.W. Are educational researchers really prepared to evaluate instructional programs? (Paper presented at the Western Regional Conference on Testing Problems. Educational Testing Service, Princeton, New Jersey, 1967.)

Skager, R.W., & Weinberg, C. Research fundamentals in educational change. Glencoe, Ill.: Scott-Foresman, (In press).

Stake, R.E. The countenance of educational evaluation. Teacher's College Record, 1967, 68, 523-540.

Stufflebeam, D.L. Evaluation as enlightenment for decision-making. (Address delivered at the Working Conference on Assessment Theory, Sarasota, Florida, January 19, 1968.) Printed by the Evaluation Center, Ohio State University.

Thorndike, R.F., & Hagen, E. Measurement and evaluation in education. New York: John Wiley and Sons, 1955.

Wallace, J.G. Concept growth and the education of the child: A survey of research on conceptualization. New York University, 1967.

Webb, J.H. The evaluation of Xerox's LI X scanner and printer as a device for use in whole item storage and retrieval. Test Development Systems Report. Princeton, New Jersey: Educational Testing Service, 1968.

Wood, D.A. Test construction. Charles E. Merrill, 1960.

This publication is published pursuant to a contract with the U.S. Office of Education, Department of Health, Education and Welfare. Points of view or opinions stated do not necessarily represent official U.S.O.E. position or policy.